Charting a Course through a Rapidly Changing Scientific Paradigm

Data—the core of any scientific endeavor—is at once a great asset and an impediment to scientific progress. Today's tools for collecting, using, and sharing data offer unprecedented opportunities for new insights and societal benefit. The amount of available data is growing exponentially, and the landscape of tools for analyzing and sharing data is evolving rapidly. These trends, in the context of a changing scientific culture, are leading to fundamental shifts in the practice of science—presenting both great opportunities and great challenges.

Earth scientists, data scientists, business leaders, and the U.S. government have begun working on many fronts to more effectively harness the power of Earth science data for collective benefit. These efforts are necessary and they are making important strides, but they are not sufficient. A unifying vision is needed to guide the development of cohesive, effective strategies and policies and address the data grand challenges that span multiple domains and organizations.

Members of the Federation of Earth science Information Partners (ESIP) and representatives from the National Research Council (NRC) have met regularly since January 2013 to discuss the possibility of a high level study to accomplish that convergence. In July 2013, a plenary discussion at the Summer ESIP meeting brought these issues into focus as panelists considered the need and feasibility of establishing an NRC study on data developments, management, and stewardship in the Earth sciences realm.

We find ourselves at a crossroads. Vast new troves of information are becoming available all the time as new sensors are deployed, networks are built, and open-source tools are shared. But we need high-level strategic guidance to effectively and efficiently channel the potential of all that data into real benefits for science and society. This paper builds on the ongoing ESIP/NRC conversation to outline the data opportunities we face, ongoing efforts in this realm, and a potential path forward.

The Vision: Riding the Data Tide

The Earth science community is well aware of the fact that the data sets available today are vastly larger, more numerous, and more complex than ever before. Recent directives aimed at making many of the U.S. government's data sources publicly accessible will likely reinforce this trend (Obama 2013; Holdren 2013). Collective investments, such as data sets and models produced by government agencies and the scientific community, hold the potential for tremendous benefits; weather and Global Positioning System (GPS) data are two examples of publicly-available government resources that have led to numerous scientific and commercial innovations. It is our shared responsibility to maximize the return on such collective investments.

Optimizing the use of data is no simple feat, however. The exponential growth of data is akin to a giant tide; as we rise to the tide's crest, we can either channel its energy into major scientific advances or drown in the flood. To effectively "ride the tide," we must confront current trends and guide future developments in three main realms: big data, new computational tools, and changes in

the practice of science.

Harnessing Big Data

Sensors are more powerful and ubiquitous than ever. Satellites are cheaper. Smartphones, tablets, and computers provide constant connectivity. Drones proliferate. Robots improve. The common thread running through these trends is data—lots of it. Enormous data collections are accumulating in government-run databases such as those under the purview of the National Aeronautics and Space Administration (NASA), National Oceanic and Atmospheric Administration (NOAA), National Science Foundation, United States Geological Survey, Department of Energy, and Environmental Protection Agency, as well as many academic and industry-run databases. For example, NASA's Global Change Master Directory "holds more than 29,000 Earth science data set and service descriptions [covering] subject areas within the earth and environmental sciences".

Some of these data sets are highly likely to be used far beyond their original point of collection and for purposes other than their original intent. Maximizing the usefulness of these collections presents challenges throughout the entire data life cycle, including planning, collection, storage, documentation, maintenance, and preservation.

Capitalizing on Computational Advances

The past decade has seen remarkable progress in the tools available for using and sharing data. Our computational and analytical capacity has exploded as processors become increasingly powerful and innovative new software is deployed. Middleware has become more important as the need for integration and distributed computing has grown. Techniques for mining and analyzing unstructured and "dark" (or ignored) data continue to grow more sophisticated.

But as computational capabilities increase and new tools are deployed, interface mismatches proliferate and the lack of interoperability among systems and software becomes more troublesome. In addition to the challenges of dealing with multiple data formats, there is a need to address the broader, fundamental issues in data representation. Such issues must be addressed from a cross-sector and interdisciplinary perspective.

Changing the Practice of Science

The evolving technological landscape has led to important new trends in the way science is done. Today's data resources are so large and complex that their utility often extends far beyond a single research project or discipline. Scientists, as a result, increasingly must navigate across fields, bridge between organizations, and work within funding and employment structures that are more fluid and adaptive than the discrete research projects and tenure-centric paradigm of the past. At the same time, science has in many ways become more open and "democratic." Offering open data, creating open-source software, and publishing on open-access platforms allows more people greater access to information and analytical tools, and the proliferation of citizen science and crowdsourcing efforts are broadening the scientific community. As a result of these trends, reproducibility has become more important. As more scientific claims are made and more results are being challenged, there has been a growing demand for greater transparency and accountability in science. One compelling vision for the future is that of an "executable publication," which would allow readers to follow links

in a scientific publication to acquire primary data and execute code to verify research results (see box).

These trends are exciting, but they do not fit well within our current siloed research structure. To maximize the benefits of new developments in the practice of science, we need a coherent, cross-domain, and cross-organizational vision for data sharing and publication.

[Box]
Data as Nanopublication
"The goal is to have a world in which all of the science literature is online, all of the science data is online, and they interoperate with each other" –Jim Gray, 2007 (Hey, et al. 2009)

The scientific literature is no longer a collection of journals crowding a library shelf. Instead, scientific contributions in the form of data, software, and scientific conclusions are now found in a variety of formats in the cloud. In a recent blog post, Adrian Giodiani shared a vision of research papers as executable data objects that turn all stages of the scientific process, including data and analysis, into a reducible and citable format (Giodiani 2013). In a similar vein, the Global Biodiversity Information Facility and Pensoft Publishers have developed a pilot workflow to publish biodiversity and ecology data. Metadata descriptions are used to automatically generate XML-based "data papers," which are then cross-linked with any analyses stemming from them. Such approaches give appropriate credit to data creators or collectors while also providing a mechanism for making data publicly and permanently available to all.

We are rapidly moving away from a linear paradigm that puts publication at the end of the scientific process toward one in which data can be shared as a "nanopublication" in the middle of the process, allowing many researchers to use it to generate insights and discoveries.
[end box]

Important Efforts to Date
Many organizations, initiatives, and advisory groups have made important strides toward solving our data challenges, including:

- The Blue Ribbon Task Force on Sustainable Digital Preservation and Access
- Data.gov
- EarthCube
- National Research Council
- NASA's Earth science Data System Working Groups
- NOAA's Environmental Data Management Committee
- Research Data Alliance
- Sustainable Digital Data Preservation and Access Network Partner (DataNet) and its funded projects, DataONE and the Data Conservancy

Concurrent with these efforts has been a growing recognition of the data science field and the value of data as a scientific contribution. The field of data science, drawing from and building on disciplines

such as information science, library science, and computer science, is maturing quickly. The Data Science Journal provides a central resource for data science trends and developments; the National Consortium for Data Science, launched in 2012, provides a forum for exchange, workforce development, and cross-sector collaboration to tackle big data science challenges. At the same time, data contributors and creators of scientific software are increasingly being rewarded for their work in the form of citable publications (such as in *Scientific Data*, an "open-access, online-only publication for descriptions of scientifically valuable datasets" launching in 2014 by the journal *Nature*).

All of these efforts—both domain-specific initiatives and broader data science efforts—are promising and necessary, but lack the guidance of an overarching strategy. Although the need for effective data management and stewardship is systemic to the practice of science across all organizations, it often is being solved piecemeal, with each organization responding to its own needs ad hoc and setting its own data standards and policies. This leads to duplication of effort and makes it difficult to coherently address trans-disciplinary and cross-sector issues. At the same time, our data management and stewardship problems continue to mount. Without robust preservation structures and methods, data can be inadvertently lost over time. Strategic guidance is needed to illuminate the path forward and optimize the use of our time, funding, and creative energy.

Charting a Path Forward

"We have a shared responsibility to create and implement strategies to realize the full potential of digital information for present and future generations." –*e*GY Declaration, 2007 (CoBabe-Ammann, et al. 2007)

Though six years have passed since the *e*GY declaration, we remain mired in the challenges of dealing with scientific data. As these issues become increasingly complex, there is growing interest in developing a unified vision to transform our data challenges into scientific opportunities. The NRC is the logical coordinator of this effort. As the operating arm of the National Academy of Sciences, the NRC has led many influential studies and has the expertise needed to convene the right leaders to guide our future efforts. An NRC-led study would identify major gaps in data management knowledge and practices, set research priorities for scientific data management and stewardship, and offer a consensus view of opportunities to retain the role of the United States as a global scientific leader.

The possibility of an NRC study to help tackle our data challenges was the topic of a panel discussion entitled "The potential value and benefits of a Data Decadal Survey" held during the Summer 2013 ESIP meeting. Panelists spanning the private, academic, and non-profit sectors stressed the need for a broad effort to identify our highest-priority scientific questions, develop an integrated strategy to build the scientific environment to address those questions, and guide future efforts in a multi-sector, multidisciplinary, and international context. A follow-on workshop planned for the winter 2014 ESIP meeting will further refine this vision.

An NRC study would offer many unique benefits. As an independent advisory organization, the NRC is a well-respected resource for informing priorities in the federal agencies, executive branch leadership, and Congress. The NRC is ideally positioned to ensure the concerns and needs of all

stakeholders—from the private sector to academic researchers to government agencies and policy makers—are heard and integrated into the overarching vision. With its longstanding role as the central forum and voice of the scientific community, the NRC is uniquely capable of drawing upon the top echelon of scientific leaders to guide visionary science and chart a path to achieving it through targeted research investments, cultural changes, and strategic coalitions.

A coordinated effort at the highest level is needed to allow the Earth sciences to fully and effectively capitalize on the unfolding data revolution. We hope you will join this conversation by contributing your needs, challenges, and ideas to our ongoing dialogue. A site for community input is under development and will be available at esipfed.org in the near future. We look forward to working together to solve our common problems and channel our creative energy into a new wave of discovery and innovation.

References

Baker, D. Barton, C.E., Peterson, W.K., & Fox, P. "Informatics and the 2007-2008 Electronic Geophysical Year." *EOS Transactions*, 89(48): 485-6.

CoBabe-Ammann, E., Peterson, W.K., Baker, D. Fox, P., & Barton, C. (2007). "The Electronic Geophysical Year (2007–2008): eScience for the 21st Century." *Geophysics*, 26:1294-5.

Giordani, A. "Scientific publishing 2.0: moving the compute to the data rather than moving the data to the computers", soapboxscience, Nature.com Blogs, Blog post February 6, 2013.

Global Change Master Directory. National Aeronautics and Space Administration. http://gcmd.gsfc.nasa.gov/learn/

Hey, T., Tansley, S., & Tolle, K., Eds. (2009). "The Fourth Paradigm: Data-Intensive Scientific Discovery." Microsoft Research, Redmond, Washington.

Holdren, J.P. "Memorandum for the Heads of Executive Departments and Agencies—Increasing Access to the Results of Federally Funded Scientific Research." February 22, 2013.

Scientific Data. Nature Publishing. http://www.nature.com/scientificdata/

Obama, B. "Executive Order—Making Open and Machine Readable the New Default for Government Information." May 9, 2013.

Acknowledgements

We greatly appreciate the panelists of the ESIP July 2013 Workshop for their time, input and support: Dan Baker, LASP, Stan Ahalt, RENCI, Todd Vision, UNC and NESCent, Michael Tiemann, RedHat. We

thank the ESIP Data Study Working Group for input, review, and ongoing support. We thank Anne Johnson, freelance writer, for her excellent editorial contributions. Finally, we thank the EOS editors and reviewers for helpful suggestions and comments.

Support for Robert Downs was provided under NASA Contract NNG13HQ04C for the Continued Operation of the Socioeconomic Data and Applications Center (SEDAC). Anne Wilson gratefully acknowledges the support provided by the Laboratory for Atmospheric and Space Physics (LASP). Ramapriyan worked on this paper as a part of his official duties as a U.S. government employee. William Michener is supported by the U.S. National Science Foundation (Grant #ACI-0830944 and #IIA-1301346). Any opinions, findings, and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of their employers or funders.